### Scope

The NCBI Conserved Domain Database (CDD) is a protein domain annotation resource, and includes a collection of ancient protein domains (> -0.5 Billion years old) and hierarchies of domains related by common descent. Each conserved domain (CD) record consists of a protein multiple sequence alignment, a consensus sequence, and a PSSM (position-specific score matrix) that quantitatively represents the information in the alignment. Whenever possible, 3D structural information is used to define and refine the alignment models. CDD is supported by an active curation effort that identifies new domains, builds domain hierarchies, and continually updates existing records with new sequences. In addition to NCBI curated entries, it also contains entries from Pfam, Smart, Protein Clusters, TIGR, and last release of COG.

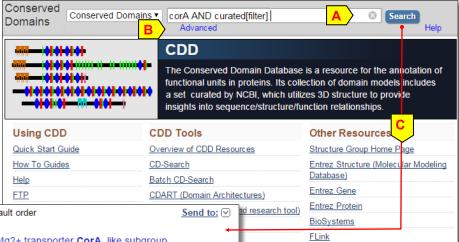
### Access to CDD data

CDD is part of the NCBI Entrez system and is extensively linked with other NCBI data. CDs can be found by direct text searching from the CDD homepage. Relevant CDDs present in a protein entry will be linked from that record. PSSMs of each CD are represented in scoremat format, and these scoremats have been made available for search with a protein query sequence through a specialized BLAST program called reverse psi-blast (RPS-BLAST). These scoremats are also available for download from the ftp site so that a selected subset can be used to build custom domain databases for local RPS-BLAST searches using the standalone BLAST+ package.

Direct text query of the CDD database: Indirect query with a sequence using RPS-BLAST: RPS-BLAST access from BLAST Homepage: Download of scoremat data files by FTP: https://www.ncbi.nlm.nih.gov/cdd/ https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi https://blast.ncbi.nlm.nih.gov/ ftp://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/

# **Searching with Text Query**

Entering a set of query terms in the search box and pressing the "Search" button (A) from the CDD homepage retrieves relevant records from the database. The "Advanced" page (B) provides access to indexing fields and their terms, which can be used to construct complex query terms for more specific retrieval.





Searching with (corA) AND "curated"[filter] retrieves group of conserved domain records shown in summary format (C). These domains are present in well-conserved family of genes commonly found in bacteria and archaea, that encode a bi-valent cation transporter for Mg<sup>2+</sup> influx or Zn<sup>2+</sup> efflux. The thumbnail (D) indicates that the domain contains a structure record and can provide additional information on sequence-structure relationship. Clicking the title (E) opens the full record.

Page 2 **CDD: Conserved Domain Database** 

### Elements of a CDD Record

A CDD records contains several elements that provide key information for the domain.

- Sequence alignment: Each aligned sequence is matched to a record in Entrez Protein database. If sequences with structural data (from the PDB database) exist, one of them will be chosen as the mater sequence of the
- Consensus sequence: Each position in the consensus sequence contains the residue with the highest frequency in that column of the alignment. For a column to be included in the consensus sequence, at least 50% of the sequences must have an aligned residue in that column.
- PSSM: For each position in the consensus, frequency ratios (expected/observed) are calculated for each amino acid, and these frequencies are converted to scores in a PSSM. The PSSM thus has the same number of columns as the consensus and precisely defines the extent of the domain.

# A Sample CDD Record

In the detailed record view of a conserved domain record (shown at the right), the summary on the domain function and taxonomic distribution (A) is displayed at the top. Clicking the thumbnail (B) opens the protein structure in Cn3D to allow interactive examination of the sequence alignment in the context of the structure and its feature annotation. Portlets in the left column provide statistics and relevant links to other tools, such as PSSM Viewer and CD Tree. An embedded display of CD Tree is also shown to the right (C). The structure guided functional annotation of this domain is summarized in the "Conserved Features/Sites" section (D). Features and sites annotated are listed in individual tabs. The protein sequence alignment is shown at the end (E), where the color of the residuals indicates the level of conservation and "#" and yellow column background indicate a conserved feature selected

Sequence Alignment

Feature 1

gi 257170536 gi 353194386

gi 307611750

gi 257797392

gi 88187722

gi 110620367

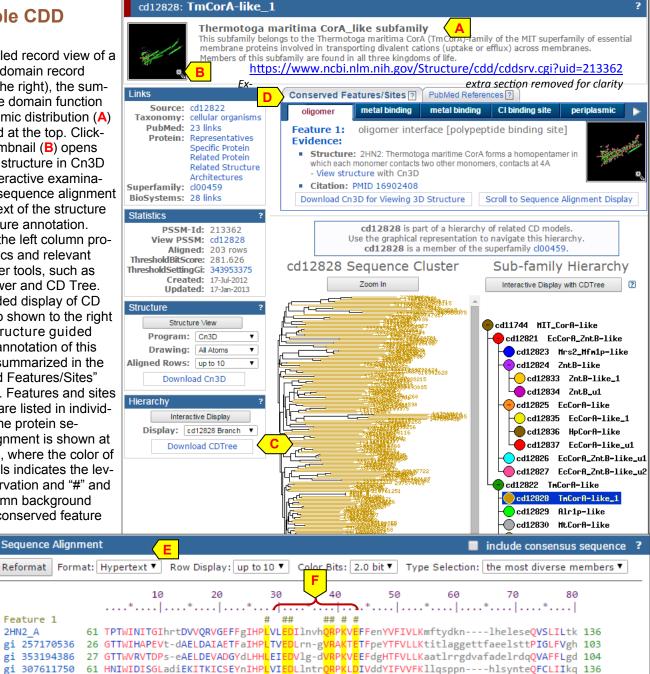
2649821

gi 282157735

307156401

2HN2 A

(**F**).



gi

gi

65 GSKWVIITGLantQLINDICEFYgMHPLTTEDIlnthQRPKIESFddYIYIVIKvilpmed----egiyseQASIILfk 139

56 WNHWIQCIGVhdpAVLEGFRKHFgIHPLALEDVmnteHRPKLEHFpdSLFLVCKwlgfned----rslsdnHISLIMqn 130 66 GICWVHISGLKdiDKINQILSDAILHP<mark>L</mark>IA<mark>ED</mark>Vfntk<mark>GRAKVE</mark>DFgdTLFIILDiilsng-----telqsrKINIITke 139

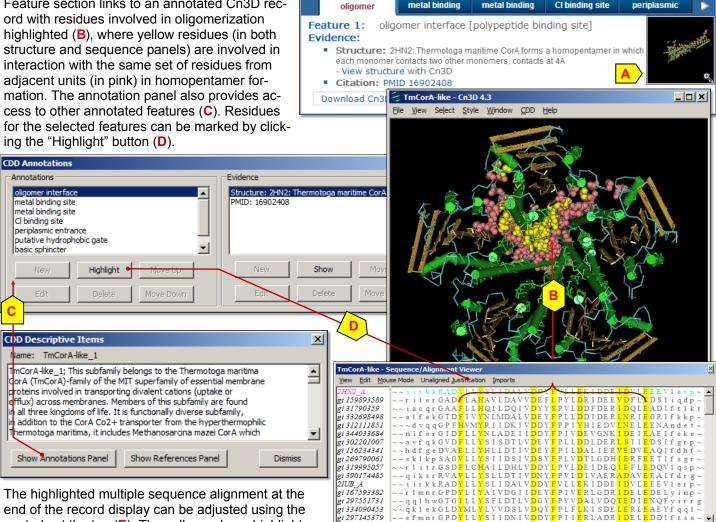
27 YNRWISLTRPe-kQDIDLVVSKFsLHPLVIEDItnprEIPKVDEYagYTFVVTDipeien-----esviihKLYIILgk 99

63 KKLWIDVVGVhdeSLIAKICEFLgIHPLAAEOIlntaQRVKIEDYddHLFLVLKillyne-----tleidQLSLVLkk 135

27 CNKWVSMLSPs-pQEFKAVADAFeLHPLVVEDMandkELPKVNEYahYTFLILDvpehdd-----efaiskLYIVIgr 98

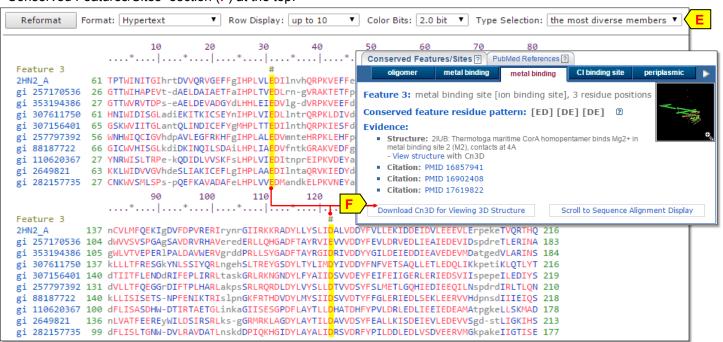
# A Sample CDD Record (cont.)

Furthermore, the structure thumbnail (A) in the Feature section links to an annotated Cn3D record with residues involved in oligomerization highlighted (B), where yellow residues (in both structure and sequence panels) are involved in interaction with the same set of residues from adjacent units (in pink) in homopentamer formation. The annotation panel also provides access to other annotated features (C). Residues for the selected features can be marked by clicking the "Highlight" button (D).



Conserved Features/Sites ? | PubMed References ?

end of the record display can be adjusted using the controls at the top (E). The yellow column highlight pattern will be updated to reflect selection in the "Conserved Features/Sites" section (F) at the top.



gi 219541761

### **Related Software Tools**

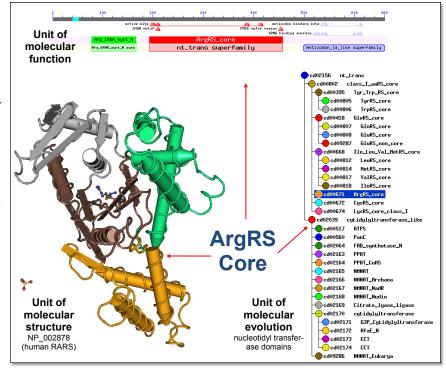
A set of tools are available for working with CDD related data.

- CDTree: analyzes sequence alignments, constructs phylogenetic trees, and CD hierarchies; reads/writes CDs.
- Cn3D 4.3: renders and aligns structures; creates and edits sequence alignments; reads/writes CDs; reads mFASTA, writes gapped FASTA, and PSSMs as NCBI scoremats.
- makeprofiledb: command-line utilities (part of the BLAST package) that converts scoremats into a rpsblast database.
- **psiblast:** command-line program that creates PSSMs iteratively from proteins that are sequence-similar to a query protein; writes frequency scoremats only; can initiate a search with a frequency scoremat.
- fa2cd.exe: a command line utilities that converts gapped FASTA into a cd file readable by CD Tree and Cn3D.
- **rpsblast**: searches a input protein query against a CDD database generated by makeprofiled to identify domains present in the query. Both standalone version (as part of the blast+ package) and a web service are available. All protein BLAST searches with have rpsblast searches performed simultaneously.
- Batch-cd search: a web service that takes a set of protein sequences and performs CDD searches with rpsblast.

# CDD Summary and Annotation Files Available by FTP

Extensive data files are available from the cdd ftp site. Representative entries or directories are listed below.

- Full CDs (acd.tar.gz): these files, with .acd file extension, contain all data for each CD record. These files can be read and written by both Cn3D and CDTree.
- CD scoremates (cdd.tar.gz): these files are ASN.1 representation of the PSSMs containing only final scores, and can be used as input to formatrpsdb. They cannot be used to initiate a PSI-BLAST search.
- FASTA alignment (fasta.tar.gz): each of these files contains the entire CD alignment, including the consensus sequence, in gapped FASTA format. These files are suitable for importing CD alignments into sequence analysis software.
- CD Summaries (cddid.tbl.gz): this file contains the PSSM-id, CD accession, title, description and PSSM length (number of columns) of each CD.



- Curated Annotations (cddannot.dat.gz): this file is an index of all annotations present on NCBI curated CDs.
- Master sequences (cddmasters.fa.gz): this file contains the master sequence of each CD in FASTA format.
- **CDD versions (cdd.versions)**: this file contains accessions, PSSM-ids, names, and created dates of all current and previous versions of CDD records.
- **Preformatted CDD database** (big\_endian and little\_endian): these directories contain preformatted database files for use with rpsblast in a local installation.

#### References

- Marchler-Bauer A, et al. (2013) CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 41: D348-52. Epub 2012 Nov 28. [PMID: 23197659]
- 2. Marchler-Bauer A, Bryant SH (2004). CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32: W327-31. [PMID: 15215404]
- Marchler-Bauer A, et al. (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35: D237-40. [PMID: 17135202]
- 4. Geer LY, et al.. (2002) CDART: protein homology by domain architecture. Genome Res. 12(10): 1619-23. [PMID: 12368255]